Report of project 2

Algorithm	Overall	MAE of test 5	MAE of test 10	MAE of test 20
Cosine	0.793	0.829	0.774	0.776
Pearson	0.813	0.883	0.794	0.769
Pearson with IUF	0.807	0.871	0.788	0.770
Pearson with case modification	0.784	0.816	0.782	0.761
Pearson with both IUF and modification	0.854	0.909	0.841	0.820
Item_based Algorithm	0.856	0.913	0.84	0.820
My own algorithm	0.778	0.803	0.763	0.767

For all of the algorithms above, I tried different Ks and finally I chose to set K to 150,which made my best result. My own algorithm is just the linear combination of cosine similarity and pearson correlation,given the weight of 0.6 on Cosine and 0.4 on Pearson correlation. I also implemented Dirichlet smoothing to make a much better result.

For the results, I think most of them are quite reasonable because they are all in the range around 0.80. But the result of Pearson correlation with both IUF and case modification and item_based algorithm are not that good. I think it may be due to overfitting or excessive complexity.

Advantages and disadvantages:
1.Cosine similarity:
Advantage:easy to implement and good to handle high_dim sparse data
Disadvantage: do not account for the magnitude of the vectors;sensitive to 0 vectors
2.Pearson correlation:
Advantage: account for the linear relationship between different variables
Disadvantage: sensitive to outliers
3.Pearson correlation with IUF:
Advantage: adjust the frequency of user's rating,given less weight to users
Disadvantage:increase the complexity and may not always give a better result

4.Pearson correlation with case modification:
Advantage: enhance similarity measure
Disadvantage: computational complexity
5.Pearson correlation with both IUF and case modification:
Advantage: combine IUF with case modification,improve the result
Disadvantage:increase the complexity and may lead to overfit problem which can give a not good result
6.Item_based:
Advantage:more stable and can lead to more consistent recommendations.
Disadvantage:if the item number is large, there will be a huge computational intensive problem.
7.my own algorithm:
Advantage: given weight to different methods to generate a better result
Disadvantage:long running time.